CLAIMS:

1. A method for clustering data points with defined quantified relationships between them comprising the steps of:
   - obtaining lead value for each data point either by deriving from said quantified relationships or as given input,
   - ranking each data point in a lead value sequence list in descending order of lead value,
   - assigning the first data point in said lead value sequence list as the leader of the first cluster, and
   - considering each subsequent data point in said lead value sequence list as a leader of a new cluster if its relationship with the leaders of each of the previous clusters is less than a defined threshold value or as a member of one or more clusters where its relationship with the cluster leader is more than or equal to said threshold value.

2. The method as claimed in claim 1, wherein said relationships between data points are symmetric or asymmetric.

3. The method as claimed in claim 1, wherein the lead value of each data point is determined by taking the sum of relation values of each of the other data points to said data point.

4. The method as claimed in claim 1, wherein said threshold value is adaptively found for a given number of clusters.

5. A method for organizing a set of data points into a hierarchy of clusters wherein the method claimed in claim 1 is first used to cluster the data points into sets of small sizes, each smaller set is further subclustered using the method and subclustering is repeated until a terminating condition is reached.

6. The method as claimed in claim 1 applied to text summarization of a single document or a collection of documents comprising the steps of:

- segmenting the given input text into blocks such as sentences, a collection of sentences, paragraphs,
- excluding words belonging to a defined list of 'stop' words,
- replacing words by their unique synonymous word, if it exists, from a given collection of synonyms,
- application of stemming algorithms for mapping words to root words,
- representing the resulting blocks of text, with respect to a dictionary which is either given or computed from the input text, by a binary vector of size equal to the number of words in the dictionary whose *i*th element is 1 if *i*th word in the dictionary is present in the block,
- computing the relationship between any data points di and dj by evaluating R(di,dj) = |dj.Tdi|/|dj| wherein T is a thesaurus matrix whose *ij*th element reflects the extent of inclusion of meaning of *j*th word in the meaning of *i*th word, and
- clustering the data points wherein the lead value of each data point is determined by taking the sum of relation values of each of the other data points to said data point, the threshold value is adaptively found for a given number of clusters and the set of leaders of the resulting clusters summarize the given text.

7. The method as claimed in claim 6 wherein said dictionary is computed by taking the fraction of words, excluding the stop words, with highest tfidf value, which is given by:

tfidf(*wi*) = *tfi* * log(*N* / *dfi*)

where tfidf(*wi*) is the lead value of data point *wi*, *tfi* = the number of times the data point *wi* occurred in the whole text, *dfi* = the number of documents containing the data point *wi* and *N* = the total number of documents in the text.

8. The method as claimed in claim 6 wherein said thesaurus matrix is either a given, identity matrix or computed from a collection of documents.

9. The method as claimed in claim 6 wherein each block is represented by a vector whose *i*th element represents the frequency of occurrence of *i*th word in the block.

10. A method for organizing a set of text documents into a hierarchy of clusters wherein the method claimed in claim 6 is first used to cluster the given documents into sets of small sizes, each smaller set is further subclustered using the method and subclustering is repeated until a terminating condition is reached.

11. The method as claimed in claim 10 applied to organize the results returned by any information retrieval system in response to an user query into an hierarchy of clusters.

12. The method as claimed in claim 11, wherein the hierarchy is used to aid the user in modifying his/her query and/or in browsing through the results.

13. The method as claimed in claim 11, wherein the information retrieval system is any search engine retrieving Web documents.

14. The method as claimed in claim 5, applied to vocabulary organization for a group of documents wherein the data points are the words in the dictionary of the vocabulary, the lead value of a word is either its frequency of occurrence in the collection, the number of documents containing the word or its tfidf value, the relationship *R(di,dj)* denotes the fraction of documents containing the *j*th word that also contain *i*th word, and the clustering produced by the application of the method results in a structured hierarchical organization of the vocabulary.

15. The method as claimed in claim 14, wherein the structured vocabulary is used to provide text summarization for the associated documents.

16. The method as claimed in claim 14 applied to customer profiling wherein the dictionary is built and the vocabulary is organized using the documents that are viewed by the customer.

17. The method as claimed in claim 5 wherein data points correspond to the products cataloged in the store, the lead value of a product is its per unit profit, its per unit value or the number of items sold per unit time, and the relationship between the products is either explicitly defined or derived from the purchase data.

18. The method as claimed in claim 17 wherein the product di is related to the product dj by the fraction of customer transactions containing dj that also contain di.

19. The method as claimed in 17 applied to analyze sales of a store for the merchant or to organize the layout of the store to facilitate easy access to products.

20. The method as claimed in 17 applied to personalize the electronic store layout to an individual customer by using the relationship that is specific to the customer.

21. The method as claimed in claim 5, applied to customer segmentation for a sales or service organization wherein the data points are the customers in the data base, the lead values are their total purchase amount per unit time, their income, the number of times customers visited the store, or the number items bought by the customer, the relationship between customers is either explicitly defined or derived from some relevant data, with the resulting clustering reflecting a structured grouping of customers with similar performances.

22. The method as claimed in claim 21, wherein the customer di is related to the customer dj by the fraction of products bought by dj that are also bought by di.

23. A system for clustering data points with defined quantified relationships between them comprising:

- means for obtaining lead value for each data point either by deriving from said quantified relationships or as given input,
- means for ranking each data point in a lead value sequence list in descending order of lead value,
- means for assigning the first data point in said lead value sequence list as the leader of the first cluster, and

- means for considering each subsequent data point in said lead value sequence list as a leader of a new cluster if its relationship with the leaders of each of the previous clusters is less than a defined threshold value or as a member of one or more clusters where its relationship with the cluster leader is more than or equal to said threshold value.

24. The system as claimed in claim 23, wherein said relationships between data points are symmetric or asymmetric.

25. The system as claimed in claim 23, wherein the means for obtaining lead value of each data point is by taking the sum of relation values of each of the other data points to said data point.

26. The system as claimed in claim 23, wherein said threshold value is adaptively found for a given number of clusters.

27. The system for organizing a set of data points into a hierarchy of clusters wherein the system claimed in claim 23 is first used to cluster the data points into sets of small sizes, each smaller set is further subclustered using the system and subclustering is repeated until a terminating condition is reached.

28. The system as claimed in claim 23 used for text summarization of a single document or a collection of documents comprising:

- means for segmenting the given input text into blocks such as sentences, a collection of sentences, paragraphs,
- means for excluding words belonging to a defined list of 'stop' words,
- means for replacing words by their unique synonymous word, if it exists, from a given collection of synonyms,
- means for applying stemming algorithms for mapping words to root words,
- means for representing the resulting blocks of text, with respect to a dictionary which is either given or computed from the input text, by a binary vector of size equal to the number of words in the dictionary whose *i*th element is 1 if *i*th word in the dictionary is present in the block,

- means for computing the relationship between any data points di and dj by evaluating R(di,dj) = |dj.Tdi|/|dj| wherein T is a thesaurus matrix whose *ij*th element reflects the extent of inclusion of meaning of *j*th word in the meaning of *i*th word, and
- means for clustering the data points wherein the lead value of each data point is determined by taking the sum of relation values of each of the other data points to said data point, the threshold value is adaptively found for a given number of clusters and the set of leaders of the resulting clusters summarize the given text.

29. The system as claimed in claim 28 wherein said dictionary is computed by taking the fraction of words, excluding the stop words, with highest tfidf value, which is given by means of:
tfidf(*wi*) = *tfi* * log(*N* / *dfi*)
where tfidf(*wi*) is the lead value of data point *wi*, *tfi* = the number of times the data point *wi* occurred in the whole text, *dfi* = the number of documents containing the data point *wi* and *N* = the total number of documents in the text.

30. The system as claimed in claim 28 wherein said thesaurus matrix is either a given identity matrix or computed from a collection of documents.

31. The system as claimed in claim 28 wherein each block is represented by a vector means whose *i*th element represents the frequency of occurrence of *i*th word in the block.

32. A system for organizing a set of text documents into a hierarchy of clusters wherein the system claimed in claim 28 is first used to cluster the given documents into sets of small sizes, each smaller set is further subclustered using the system and the subclustering is repeated until a terminating condition is reached.

33. The system as claimed in claim 32 used to organize the results returned by any information retrieval system in response to an user query into an hierarchy of clusters.

34. The system as claimed in claim 33, wherein the hierarchy of clusters is used to aid the user in modifying his/her query and/or in browsing through the results.

35. The system as claimed in claim 33, wherein the information retrieval system is any search engine retrieving Web documents.

36. The system as claimed in claim 27, used for vocabulary organization for a group of documents wherein the data points are the words in the dictionary of the vocabulary, the lead value of a word is either its frequency of occurrence in the collection, the number of documents containing the word or its tfidf value, the relationship *R(di,dj)* denote the fraction of documents containing the *j*th word that also contain *i*th word, and the clustering produced by the system results in a structured hierarchical organization of the vocabulary.

37. The system as claimed in claim 36, wherein the structured vocabulary organization is used to provide text summarization for the associated documents.

38. The system as claimed in claim 36 used for customer profiling wherein the dictionary is built and the vocabulary is organized using the documents that are viewed by the customer.

39. The system as claimed in claim 27 wherein data points correspond to the products cataloged in the store, the lead value of a product is its per unit profit, its per unit value or the number of items sold per unit time, the relationship between the products is either explicitly defined or derived from the purchase data.

40. The system as claimed in claim 39 wherein the product di is related to the product dj by the fraction of customer transactions containing dj that also contain di.

41. The system as claimed in claim 39 used for analyzing sales of a store for the merchant or for organizing the layout of the store to facilitate easy access to products.

42. The system as claimed in 39 used to personalize the electronic store layout to an individual customer by using the relationship that is specific to the customer.

43. The system as claimed in claim 27, used for customer segmentation for a sales or service organization wherein the data points are the customers in the data base, the lead values are their total purchase amount per unit time, their income, the number of times customers visited the store, or the number items bought by the customer, the relationship between customers is either explicitly defined or derived from some relevant data, with the resulting clustering reflecting a structured grouping of customers with similar performances.

44. The system as claimed in claim 43, wherein the customer di is related to the customer dj by the fraction of products bought by dj that are also bought by di.

45. A computer program product comprising computer readable program code stored on computer readable storage medium embodied therein for clustering data points with defined quantified relationships between them, comprising:

- computer readable program code means configured for obtaining lead value for each data point either by deriving from said quantified relationships or as given input,
- computer readable program code means configured for ranking each data point in a lead value sequence list in descending order of lead value,
- computer readable program code means configured for assigning the first data point in said lead value sequence list as the leader of the first cluster, and
- computer readable program code means configured for considering each subsequent data point in said lead value sequence list as a leader of a new cluster if its relationship with the leaders of each of the previous clusters is less than a defined threshold value or as a member of one or more clusters where its relationship with the cluster leader is more than or equal to said threshold value.

46. The computer program product as claimed in claim 45, wherein said relationships between data points are symmetric or asymmetric.

47. The computer program product as claimed in claim 45, wherein said computer readable program code means configured for obtaining lead value of each data point is by taking the sum of relation values of each of the other data points to said data point.

48. The computer program product as claimed in claim 45, wherein said threshold value is adaptively found for a given number of clusters.

49. A computer program product for organizing a set of data points into an hierarchy of clusters wherein the computer program product claimed in claim 45 is first used to cluster the data points into sets of small sizes, each smaller set is further subclustered using the computer program product and the subclustering is repeated until a terminating condition is reached.

50. The computer program product as claimed in claim 45 configured for text summarization of a single document or a collection of documents comprising:

- computer readable program code means configured for segmenting the given input text into blocks such as sentences, a collection of sentences, paragraphs,
- computer readable program code means configured for excluding words belonging to a defined list of 'stop' words,
- computer readable program code means configured for replacing words by their unique synonymous word, if it exists, from a given a collection of synonyms,
- computer readable program code means configured for applying stemming algorithms for mapping words to root words,
- computer readable program code means configured for representing the resulting blocks of text, with respect to a dictionary which is either given or computed from the input text, by a binary vector of size equal to the number of words in the dictionary whose ith element is 1 if ith word in the dictionary is present in the block,
- computer readable program code means configured for computing the relationship between any data points di and dj by evaluating R(di,dj) =

|dj.Tdi|/|dj| wherein T is a thesaurus matrix whose *ij*th element reflects the extent of inclusion of meaning of *j*th word in the meaning of *i*th word, and

- computer readable program code means configured for clustering the data points wherein the lead value of each data point is determined by taking the sum of relation values of each of the other data points to said data point, the threshold value is adaptively found for a given number of clusters and the set of leaders of the resulting clusters summarize the given text.

51. The computer program product as claimed in claim 50 wherein said dictionary is computed by taking the fraction of words, excluding the stop words, with highest tfidf value which is given by:
tfidf(*wi*) = *tfi* * log(*N* / *dfi*)
where tfidf(*wi*) is the lead value of data point *wi*, *tfi* = the number of times the data point *wi* occurred in the whole text, *dfi* = the number of documents containing the data point *wi* and *N* = the total number of documents in the text.

52. The computer program product as claimed in claim 50 wherein said thesaurus matrix is either a given identity matrix or computed from a collection of documents.

53. The computer program product as claimed in claim 50 wherein each block is represented by a vector computer readable program code means, whose *i*th element represent the frequency of occurrence of *i*th word in the block.

54. The computer program product for organizing a set of text documents into a hierarchy of clusters wherein the computer program product claimed in claim 50 is first used to cluster the given documents into sets of small sizes, each smaller set is further subclustered using the computer program product and the subclustering is repeated until a terminating condition is reached.

55. The computer program product as claimed in claim 54 configured for organizing the results returned by any information retrieval system in response to an user query into an hierarchy of clusters.

56. The computer program product as claimed in claim 55, wherein the hierarchy of clusters is used to aid the user in modifying his/her query and/or in browsing through the results.

57. The computer program product as claimed in claim 55, wherein the information retrieval system is any search engine retrieving Web documents.

58. The computer program product as claimed in claim 49, configured for vocabulary organization for a group of documents wherein the data points are the words in the dictionary of the vocabulary, the lead value of a word is either its frequency of occurrence in the collection, the number of documents containing the word or its tfidf value, the relationship *R(di,dj)* denote the fraction of documents containing the *j*th word that also contain *i*th word, and the clustering produced by the computer readable program code means results in a structured hierarchical organization of the vocabulary.

59. The computer program product as claimed in claim 58, wherein the structured vocabulary organization is used to provide text summarization for the associated documents.

60. The computer program product as claimed in claim 58 configured for customer profiling wherein the dictionary is built and the vocabulary is organized using the documents that viewed by the customer.

61. The computer program product as claimed in claim 49 wherein data points correspond to the products cataloged in the store, the lead value of a product is its per unit profit, its per unit value or the number of items sold per unit time, the relationship between the products is either explicitly defined or derived from the purchase data.

62. The computer program product as claimed in claim 61 wherein the product di is related to the product dj by the fraction of customer transactions containing dj that also contain di.

63. The computer program product as claimed in claim 61 configured for analyzing sales of a store for the merchant or for organizing the layout of the store to facilitate easy access to products.

64. The computer program product as claimed in 61 configured for personalizing the electronic store layout to an individual customer by using the relationship that is specific to the customer.

65. The computer program product as claimed in claim 49, configured for customer segmentation for a sales or service organization wherein the data points are the customers in the data base, the lead values are their total purchase amount per unit time, their income, the number of times customers visited the store, or the number items bought by the customer, the relationships between customers is either explicitly defined or derived from some relevant data, with the resulting clustering reflecting a structured grouping of customers with similar performances.

66. The computer program product as claimed in claim 65, wherein the customer di is related to the customer dj by the fraction of products bought by dj that are also bought by di.